基于图表征知识蒸馏的图像分类方法

杨传广[1],陈路明[2],赵二虎[1],安竹林[1],徐勇军[1]
(1. 中国科学院计算技术研究所,北京 100190;2. 93114 部队,北京 100080)

摘 要: 知识蒸馏的核心思想是利用1个作为教师网络的大型模型来指导1个作为学生网络的小型模型,提升学生网络在图像分类任务上的性能. 现有知识蒸馏方法通常从单一的输入样本中提取类别概率或特征信息作为知识,并没有对样本间关系进行建模,造成网络的表征学习能力下降. 为解决此问题,本文引入图卷积神经网络,将输入样本集视为图结点构建关系图,图中的每个样本都可以聚合其他样本信息,提升样本的表征能力. 本文从图结点和图关系2个角度构建图表征知识蒸馏误差,利用元学习引导学生网络自适应学习教师网络更佳的图表征,提升学生网络的图建模能力. 相比于基线方法,本文提出的图表征知识蒸馏方法在加拿大高等研究院(Canadian Institute For Advanced Research,CIFAR)发布的100种分类数据集上提升了3.70%的分类准确率,表明本文方法引导学生网络学习到了更具有判别性的特征空间,提升了图像分类能力.

关键词: 知识蒸馏;图卷积神经网络;图像分类;元学习;表征学习
基金项目: 国家自然科学基金(No.62072434);北京市自然科学基金(No.4212027)
中图分类号: TP391.42 **文献标识码:** A **文章编号:** 0372-2112(2024)10-3435-13
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20230976

Graph-Based Representation Knowledge Distillation for Image Classification

YANG Chuan-guang[1], CHEN Lu-ming[2], ZHAO Er-hu[1], AN Zhu-lin[1], XU Yong-jun[1]
(1. *Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190, *China*;
2. *Unit* 93114 *of PLA, Beijing* 100080, *China*)

Abstract: The core idea of knowledge distillation is to use a large model as the teacher network to guide a small model as the student network, improving the performance of the student network in image classification tasks. Existing knowledge distillation methods often extract category probability or feature information as knowledge from a single input sample. They could not model the relationships between samples, decreasing the network's representation learning ability. To solve this problem, this paper introduces a graph convolutional neural network, which treats the input sample set as graph nodes to construct a relationship graph. Each sample in the graph could aggregate information from other samples, improving its own representation ability. This paper constructs the distillation loss of graph representation knowledge from the perspectives of graph nodes and relationships. It uses meta-learning to guide the student network to adaptively learn better graph representations from a teacher network, thereby improving the graph modeling ability of the student network. Compared to the baseline method, the graph-based representation knowledge distillation method improves the classification accuracy by 3.70% on the 100-classification dataset published by Canadian Institute For Advanced Research. The result indicates that the proposed method makes the student network learn a more discriminative feature space, thereby improving its image classification ability.

Key words: knowledge distillation; graph convolutional neural network; image classification; meta-learning; representation learning
Foundation Item(s): National Natural Science Foundation of China (No.62072434); Beijing Natural Science Foundation (No.4212027)

收稿日期:2023-10-19;修回日期:2023-12-19;责任编辑:王天慧

1 引言

近年来,深度神经网络在图像分类任务上获得了优越性能[1],例如,残差网络(Residual Network,ResNet)-152[2]在大规模图像分类任务上超越了人类的识别表现,但这些高性能的网络通常具有较高复杂度,难以满足实时推理的需求. 为减少模型的复杂度,模型压缩技术应运而生,主要包括结构搜索[3]、剪枝量化[4]和知识蒸馏[5]. 知识蒸馏作为由图灵奖得主Hinton[6]倡导的一项技术,被广泛用于模型压缩和性能提升. 该技术的核心思想是将一个大型网络作为教师来指导一个作为学生的小型模型,使小型网络可在不改变结构的前提下提升性能.

最初的知识蒸馏方法[6]对齐教师-学生网络之间类别的概率分布,使学生网络可以直接学习到教师网络对图像预测的答案. 虽然基于类别概率的方法通常可有效提升学生网络的分类性能,但该类型的知识仅考虑了网络的输出概率信息,忽略了网络中间层的特征信息. 后来的工作继续尝试应用特征图[7]及从中提取的信息作为蒸馏的知识类型,例如,注意力机制图[8]、互信息熵[9]和对比学习表征[10]. 神经网络中间层特征编码了从输入到输出的归纳偏置,引导学生网络学习教师网络的特征信息,可以学习到教师网络对于答案的推理过程,提升分类性能.

先前的知识蒸馏方法通常从单一的样本中提取类别概率或特征信息作为知识,没有考虑到样本之间的关联信息. 虽然之后的蒸馏方法[11]尝试将样本之间的相关性作为迁移的基本信息,但并没有利用关系信息对样本特征做进一步建模,降低了特征学习的效果. 为解决这个问题,本文提出基于图表征的知识蒸馏方法来提升学生网络性能. 本文将每一张输入图像编码为一个图结点,并使用余弦相似度计算图结点关系,对给定样本集建立关系图. 在关系图中,本文应用图卷积神经网络[12]聚合邻居结点有价值的信息,使每个图结点都可捕捉到全局信息依赖,建立更具有判别性的特征空间.

基于图表征的知识蒸馏包括基于最小平方误差的图特征蒸馏、基于欧氏距离的图关系蒸馏和基于三元组夹角的图关系蒸馏. 第一个误差项是基于最小平方对齐教师-学生网络之间的图结点特征,使学生网络的图表征能与教师网络的图表征更相似. 由于学生网络的学习能力通常比教师网络差,因此,仅仅对齐教师-学生网络的图表征可能会造成过拟合现象. 本文进一步提出图关系蒸馏,并从欧氏距离和三元组夹角2个角度来建模图关系,然后,引导学生网络去学习教师网络对应的图关系,提升表征能力.

为增强中间层的特征表达,本文在网络的中间层插入了多个辅助分支进行特征提取,然后,将辅助分支的输出特征进行图表征蒸馏. 教师-学生网络通常具有多个辅助分支,由于网络语义抽象层级不同,简单地按照层深度一对一进行匹配蒸馏可能会造成语义不一致问题[7]. 本文引入元学习网络来生成教师-学生网络之间的层匹配权重,完成自适应的层匹配蒸馏. 学生网络和元学习网络采用交替优化形式,通过数据驱动捕捉到匹配层之间的语义相关性,给语义相似的层分配更大权重,反之给语义不一致的层分配更小权重.

本文将提出的图表征蒸馏(Graph-based Representation Knowledge Distillation,GRKD)应用在多种教师-学生网络结构组合上验证效果. 在加拿大高等研究院发布的CIFAR-100数据集[13]上,GRKD相比于基线和先前最好的解耦知识蒸馏(Decoupled Knowledge Distillation,DKD)[14]分别提升了3.70%和1.60%的分类准确率. 将CIFAR-100数据集上预训练的模型迁移到下游自主学习(Self-Taught Learning,STL)-10[15]的数据集上,GRKD相比于基线和DKD分别提升3.76%和2.20%的分类准确率. 实验结果表明GRKD训练的学生网络不仅能在目标数据集上获得很好性能,也能有效迁移到未知的STL-10图像分类任务中. 图像网络(ImageNet)数据集[16]进一步验证了GRKD可以扩展到大规模的图像分类任务,分别凭借1.40%和0.57%的准确率增益超越了基线和DKD方法.

2 相关研究

2.1 图神经网络

近年来,图神经网络已被广泛应用于图数据的信息推理,包括社交网络[17]、引用网络[18]和路由预测[19]. 受卷积网络启发,图卷积神经网络被提出来进行邻居结点聚合,并被用于半监督的图结点分类任务[12]. 图卷积的思想也被迁移到计算机视觉中. Wang等人[20]应用图卷积在视频识别任务中捕捉目标之间的关系. Hu等人[21]提出一种动态的图卷积结构建模语义分割任务中不同像素的依赖关系. 袁冠等人[22]采用时空图神经网络建模手势空间连接关系和手势长距离依赖,提升手势识别性能. 与先前工作不同,本文将单张图像视为图结点的基本对象,并应用图卷积网络对每一张图像的特征进行全局信息聚合,提升图结点表征能力和分类效果.

2.2 知识蒸馏

Hinton等人[6]于2015年提出了知识蒸馏范式,利用大型模型的知识类别概率迁移给小型模型,提升小型模型性能. 最初的知识蒸馏方法仅从类别概率层面进行蒸馏,虽然学生网络能学习到教师网络的最终输

出,却没有充分利用教师网络的中间层特征信息. Romero等人[7]引导学生-教师网络特征的逐层逼近,使学生网络能够学习到教师网络对输入图像的特征推理过程. 之后的工作对中层特征进行加工得到更有信息量的知识. Zagoruyko等人[8]将特征图编码为注意力机制图重要的语义信息,剔除冗余信息. Ahn等人[9]建模交互信息熵作为连接教师-学生网络的知识纽带. Yang等人[23]提出自监督学习与知识蒸馏相结合的知识建模形式. Zhao等人[14]对知识蒸馏误差进行解耦学习,使学生网络可自适应学习教师网络信息. 近年来,知识蒸馏与对比表征学习相结合也是非常有价值的研究方向. Tian等人[10]在教师-学生网络之间应用对比表征学习来增大两者交互信息,完成隐式知识迁移. Yang等人[24]提出相互对比学习在线蒸馏多个网络. 除图像分类,对比学习与知识蒸馏也被应用于语义分割任务[25].

虽然先前知识蒸馏方法可在教师网络指导下提升学生网络的性能,但这些方法通常对单个样本进行知识提取,并没有在样本间进行知识建模,降低了学生网络的特征学习效果. 本文引入图卷积神经网络来建模样本间关系,提出图表征蒸馏,相比于先前方法有效提升学生网络在特征层面的蒸馏效果.

3　研究方法

3.1　图像分类网络结构设计

3.1.1　图像分类任务定义

1张输入图像被表达为$\boldsymbol{x}\in\mathbb{R}^{H\times W\times 3}$,其中,$H$和$W$分别代表图像的高度和宽度,3代表RGB通道. 每张图像会有1个类别标签$y\in\{1,2,\cdots,C\}$,其中,C代表图像分类任务的类别数量. 图像分类任务的目标是利用深度学习网络从C个类别中正确预测出给定图像的正确标签y.

3.1.2　标准的图像分类卷积网络结构

图像分类任务通常使用一个标准的卷积神经网络来解决,比如ResNet[2]. 卷积神经网络$f(\cdot)$通常由特征提取器$\phi(\cdot)$和线性分类器$g(\cdot)$组成. 给定1张输入图像$\boldsymbol{x}\in\mathbb{R}^{H\times W\times 3}$,卷积神经网络$f(\cdot)$将其映射为1个逻辑向量$\boldsymbol{z}=(z_1,z_2,\cdots,z_C)\in\mathbb{R}^C$,公式化为

$$\boldsymbol{z}=f(\boldsymbol{x})=g(\phi(\boldsymbol{x})) \tag{1}$$

通过使用softmax函数σ可以将逻辑向量$\boldsymbol{z}$标准化为一个类别概率分布$\boldsymbol{p}=\sigma(\boldsymbol{z})=[p_1 p_2\cdots p_C]\in\mathbb{R}^C$,如式(2):

$$p_i=\frac{\exp(z_i)}{\sum_{c=1}^{C}\exp(z_c)} \tag{2}$$

其中,$\sum_{c=1}^{C}p_c=1$,p_c代表第c个类别的概率.

3.1.3　基于辅助分支的图像分类卷积网络结构

一个标准的卷积神经网络通常逐渐对特征图进行空间维度的降采样来提取图像语义. 不同的卷积阶段会产生不同分辨率的特征图,包含了不同的表征模式[26]. 来自底层卷积阶段的高分辨率特征图通常展示了更细粒度的目标细节和高频信息,有利于小目标检测和细粒度的图像分类. 作为比较,高层卷积阶段具备更广阔的感受野,产生的低分辨率特征图信息冗余度低,包含更丰富的全局物体结构和语义特征. 联合使用多尺度的特征信息可为图像分类器提供全面分类依据,提升分类性能.

为提升卷积网络中间层特征的应用效果,本文在卷积网络的每一个降采样阶段插入一个辅助分支对特征进一步处理和优化. 卷积网络的特征提取器ϕ可被分为K个连续的阶段,即$\phi=\phi_1,\phi_2,\cdots,\phi_K$. 在每个卷积阶段$\phi_l$之后,本文插入1个辅助分支$\xi_l$,其中,$l=1,2,\cdots,K$. 每1个辅助分支$\xi_l$包括辅助特征提取器$\rho_l$和线性分类器$\pi_l$,即$\xi_l=\rho_l\circ\pi_l$,其中,$\circ$代表连接符. 辅助特征提取器$\rho_l$包括多个卷积模块来对特征进行优化.

给定输入图像$\boldsymbol{x}_i$,i代表数据集中的第i个样本,基于辅助分支的图像分类卷积网络可以产生K个特征图,表达式为

$$\boldsymbol{F}_1^{(i)}=\rho_1(\phi_1(\boldsymbol{x}^{(i)}))\in\mathbb{R}^D \tag{3}$$

$$\boldsymbol{F}_2^{(i)}=\rho_2(\phi_2(\phi_1(\boldsymbol{x}^{(i)})))\in\mathbb{R}^D \tag{4}$$

$$\cdots$$

$$\boldsymbol{F}_K^{(i)}=\rho_L(\phi_L\cdots(\phi_2(\phi_1(\boldsymbol{x}^{(i)})))\cdots)\in\mathbb{R}^D \tag{5}$$

其中,$\boldsymbol{F}_l^{(i)}$代表第l个辅助特征提取器输出的特征图,D代表特征向量的维度. 对于每个辅助特征提取器输出的特征图$\boldsymbol{F}_l^{(i)}$,对应线性分类器π_l会将其转化为类别概率分布$\boldsymbol{p}_l^{(i)}\in\mathbb{R}^C$,进一步使用图像分类的真实标签来对其进行优化.

3.2　图表征建模

给定输入特征$\boldsymbol{X}\in\mathbb{R}^{N\times D}$,其中,$N$代表结点个数,$D$代表特征向量的维度,建模得到关系图$G=\{V,E\}$,其中,$V=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_n\}(n=1,2,\cdots,N)$代表结点集合,每个结点$\boldsymbol{v}_n$是输入特征矩阵$\boldsymbol{X}$的第$n$个行向量,$\boldsymbol{v}_n\in\mathbb{R}^d$. $E=\{e^{j,k}\}\in\mathbb{R}^{N\times N}(j=1,2,\cdots,N,i=1,2,\cdots,N)$代表边集合,连接第$j$个结点$\boldsymbol{v}_j$和第$k$个结点$\boldsymbol{v}_k$的边被表达为$e^{j,k}$,其衡量了2个结点之间的关系. 本文采用余弦相似度建模2个结点之间的关系,$e^{j,k}$被公式化为

$$e^{j,k}=\frac{\boldsymbol{v}_j\cdot\boldsymbol{v}_k}{\|\boldsymbol{v}_j\|\|\boldsymbol{v}_k\|} \tag{6}$$

其中,(·)代表向量的点积,$\|\cdot\|$代表向量的模. 根据边集合可以构建图的邻接矩阵$\boldsymbol{A}\in\mathbb{R}^{N\times N}$,其中,第$j$行第$k$列的元素为$e^{j,k}$. 由于余弦相似度的对称性,邻接矩阵$\boldsymbol{A}$是对称矩阵.

相比卷积网络只能对相邻结点进行建模,图卷积网络可以利用边集使每个结点都可访问其他结点,有效聚合全局信息来增强结点特征. 给定输入结点特征$\boldsymbol{X}\in\mathbb{R}^{N\times D}$和邻接矩阵$\boldsymbol{A}\in\mathbb{R}^{N\times N}$,图卷积对每个结点建模图表征,得到输出结点特征$\boldsymbol{Y}\in\mathbb{R}^{N\times D}$:

$$\boldsymbol{Y}=M(\boldsymbol{X})=\varpi(\boldsymbol{AXW}) \tag{7}$$

其中,M代表图表征建模函数,$\boldsymbol{W}\in\mathbb{R}^{D\times D}$是可学习的权重矩阵,用于对特征做转换,$\varpi$是线性整流函数(Rectified Linear Unit, ReLU)[27]对输出特征做非线性优化. 相比于输入结点特征$\boldsymbol{X}$,输出结点特征$\boldsymbol{Y}$运用图关系加权聚合来自其他结点的特征,从而捕捉到全局的上下文依赖,得到更具有信息量的图表征.

3.3 图表征知识蒸馏

对于输入的N个样本$\boldsymbol{x}_i(i=1,2,\cdots,N)$,本文将其建模为$N$个结点. 由基于辅助分支的图像分类卷积网络结构可知,对于N个样本$\boldsymbol{x}_i(i=1,2,\cdots,N)$,会从$K$个辅助分支中产生$K$个特征向量矩阵$\boldsymbol{F}_l(l=1,2,\cdots,K)$,其中,每个第$K$层的特征向量$\boldsymbol{F}_l\in\mathbb{R}^{N\times D}$由$N$个样本第$K$层的特征向量$\boldsymbol{F}_l^{(i)}\in\mathbb{R}^{D}(i=1,2,\cdots,N)$按照行维度拼接而成. 对每个特征向量$\boldsymbol{F}_l$,本文通过图表征建模方式输出基于图的特征向量$\boldsymbol{H}_l=M(\boldsymbol{F}_l)\in\mathbb{R}^{N\times D}$,每个行向量代表每个样本结点的特征向量.

知识蒸馏通常需要教师网络f^{T}和学生网络f^{S},前者和后者分别产生基于图表征的教师特征向量$\boldsymbol{H}_l^{\mathrm{T}}(l=1,2,\cdots,K)$和学生特征向量$\boldsymbol{H}_l^{\mathrm{S}}(l=1,2,\cdots,K)$. 本文对图特征向量构造了3种蒸馏误差来训练学生网络学习教师网络中的信息,分别是:基于最小平方误差的图特征蒸馏、基于欧氏距离的图关系蒸馏、基于三元组夹角的图关系蒸馏,使学生网络学习到更好的图表征. 其中,第一种误差项属于对图结点特征的蒸馏,后两种误差项属于对图边集合的蒸馏.

3.3.1 基于最小平方误差的图特征蒸馏

给定教师特征向量$\boldsymbol{H}_l^{\mathrm{T}}\in\mathbb{R}^{N\times D}(l=1,2,\cdots,K)$和学生特征向量$\boldsymbol{H}_l^{\mathrm{S}}\in\mathbb{R}^{N\times D}(l=1,2,\cdots,K)$,本文使用最小平方误差(Minimum Squared-Error, MSE),通过回归方式使学生特征向量逼近教师特征向量. 教师网络通常具备更好的特征信息,因此,误差可直接优化得到更好的学生特征. 基于最小平方误差的图特征蒸馏公式为

$$L_{\mathrm{MSE}}\left(\boldsymbol{H}_{l^{\mathrm{T}}}^{\mathrm{T}},\boldsymbol{H}_{l^{\mathrm{S}}}^{\mathrm{S}}\right)=\frac{1}{N\times D}\sum_{n=1}^{N}\sum_{d=1}^{D}\left(\boldsymbol{H}_{l^{\mathrm{T}}}^{\mathrm{T}}[n][d]-\boldsymbol{H}_{l^{\mathrm{S}}}^{\mathrm{S}}[n][d]\right)^2 \tag{8}$$

其中,n和d分别代表特征向量矩阵的行索引和列索引;$l^{\mathrm{T}}\in\{1,2,\cdots,K\}$和$l^{\mathrm{S}}\in\{1,2,\cdots,K\}$分别代表教师网络和学生网络的辅助分支索引. 最小化误差$L_{\mathrm{MSE}}\left(\boldsymbol{H}_{l^{\mathrm{T}}}^{\mathrm{T}},\boldsymbol{H}_{l^{\mathrm{S}}}^{\mathrm{S}}\right)$用于将学生网络第$l^{\mathrm{S}}$个辅助分支产生的图特征矩阵$\boldsymbol{H}_{l^{\mathrm{S}}}^{\mathrm{S}}$逼近教师网络第$l^{\mathrm{T}}$个辅助分支产生的图特征矩阵$\boldsymbol{H}_{l^{\mathrm{T}}}^{\mathrm{T}}$,学生网络可以学到更优的图特征矩阵.

3.3.2 基于欧氏距离的图关系蒸馏

在图结点特征蒸馏之外,本文进一步考虑图关系蒸馏. 教师网络的建模图通常展现了更优的结点关系,引导学生网络学习教师网络的结点关系信息可帮助学生的图结点建立更好的图结构. 首先,采用欧式距离(Euclidean Distance, ED)对2个结点的关系进行度量. 对于第j个结点和第k个结点,其中,$j,k\in(1,2,\cdots,N)$,教师图特征和学生图特征对两者的欧氏距离分别表达为

$$d_{l^{\mathrm{T}}}^{\mathrm{T}}(j,k)=\left\|\boldsymbol{H}_{l^{\mathrm{T}}}^{\mathrm{T}}[j]-\boldsymbol{H}_{l^{\mathrm{T}}}^{\mathrm{T}}[k]\right\|_2^2 \tag{9}$$

$$d_{l^{\mathrm{S}}}^{\mathrm{S}}(j,k)=\left\|\boldsymbol{H}_{l^{\mathrm{S}}}^{\mathrm{S}}[j]-\boldsymbol{H}_{l^{\mathrm{S}}}^{\mathrm{S}}[k]\right\|_2^2 \tag{10}$$

其中,j和k都代表特征向量矩阵的行索引.

本方法引导学生网络的结点欧式距离关系去模仿对应的教师网络结点关系,使学生网络学习到更好的高阶结构化信息. 对于学生网络第l^{S}个辅助分支产生的图特征矩阵$\boldsymbol{H}_{l^{\mathrm{S}}}^{\mathrm{S}}$结点的欧氏距离去逼近教师网络第$l^{\mathrm{T}}$个辅助分支产生的图特征矩阵$\boldsymbol{H}_{l^{\mathrm{T}}}^{\mathrm{T}}$结点的欧氏距离误差公式为

$$L_{\mathrm{ED}}\left(\boldsymbol{H}_{l^{\mathrm{T}}}^{\mathrm{T}},\boldsymbol{H}_{l^{\mathrm{S}}}^{\mathrm{S}}\right)=\frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N}L_{\delta}\left(d_{l^{\mathrm{T}}}^{\mathrm{T}}(j,k),d_{l^{\mathrm{S}}}^{\mathrm{S}}(j,k)\right) \tag{11}$$

其中,L_{δ}代表平滑的范数距离,表达式为

$$L_{\delta}(a,b)=\begin{cases}\dfrac{1}{2}(a-b)^2, & |a-b|\leqslant\delta\\ \delta|a-b|-\dfrac{1}{2}\delta^2, & |a-b|>\delta\end{cases} \tag{12}$$

本文取$\delta=1$来进行计算.

3.3.3 基于三元组夹角的图关系蒸馏

前面本文使用了欧氏距离来建模图关系. 本节则采用三元组夹角的方式对图结点进行关系建模. 相比于欧氏距离的二阶建模方式,三元组夹角采用更高的三阶建模方式,因此,捕捉到更多图内部结点关系,传递更多信息. 对于第j个、第k个和第u个结点,$j,k,u\in\{1,2,\cdots,N\}$,组成的教师-学生网络图表征三元组分别为$\left(\boldsymbol{H}_{l^{\mathrm{T}}}^{\mathrm{T}}[j],\boldsymbol{H}_{l^{\mathrm{T}}}^{\mathrm{T}}[k],\boldsymbol{H}_{l^{\mathrm{T}}}^{\mathrm{T}}[u]\right)$和$\left(\boldsymbol{H}_{l^{\mathrm{S}}}^{\mathrm{S}}[j],\boldsymbol{H}_{l^{\mathrm{S}}}^{\mathrm{S}}[k],\boldsymbol{H}_{l^{\mathrm{S}}}^{\mathrm{S}}[u]\right)$,其中,$j$、$k$和$u$都代表特征向量矩阵的行索引. 三元组夹角的余弦值表达为

$$\cos\angle(j,k,u)=\boldsymbol{r}^{jk}\cdot\boldsymbol{r}^{uk} \tag{13}$$

$$\boldsymbol{r}^{jk}=\frac{\boldsymbol{H}[j]-\boldsymbol{H}[k]}{\left\|\boldsymbol{H}[j]-\boldsymbol{H}[k]\right\|_2},\ \boldsymbol{r}^{uk}=\frac{\boldsymbol{H}[u]-\boldsymbol{H}[k]}{\left\|\boldsymbol{H}[u]-\boldsymbol{H}[k]\right\|_2} \tag{14}$$

将教师-学生网络图表征三元组代入式(13)中,得到教师-学生网络三元组夹角的余弦值分别为$\cos\angle(j,k,u)^{\mathrm{T}}$和$\cos\angle(j,k,u)^{\mathrm{S}}$. 本方法利用式(12)来引导学生网络的三元组夹角去逼近对应教师网络的三元组夹角,表达为

$$L_{\cos}\left(\boldsymbol{H}_{l^{\mathrm{T}}}^{\mathrm{T}},\boldsymbol{H}_{l^{\mathrm{S}}}^{\mathrm{S}}\right)=\frac{1}{N^3}\sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{u=1}^{N}L_{\delta}\left(\cos\angle(j,k,u)^{\mathrm{T}},\cos\angle(j,k,u)^{\mathrm{S}}\right) \tag{15}$$

通过三元组夹角的图关系蒸馏,学生网络可以从教师网络中学习到更好的三阶图结点依赖关系,提升图结点的特征表达能力.

3.4　元优化图表征蒸馏

本章主要考虑怎样在基于辅助分支的教师网络和学生网络之间进行知识蒸馏层匹配(如图1所示),有效提升知识蒸馏的性能.

3.4.1　图表征蒸馏误差

教师网络的K个辅助分支和学生网络的K个辅助分支分别产生基于图表征的教师特征向量$\boldsymbol{H}_l^{\mathrm{T}}(l=1,2,\cdots,N)$和学生特征向量$\boldsymbol{H}_l^{\mathrm{S}}(l=1,2,\cdots,N)$. 文中使用$l^{\mathrm{T}}$和$l^{\mathrm{S}}$分别索引教师-学生网络的辅助分支序号,其中,$l^{\mathrm{T}}\in\{1,2,\cdots,K\}$和$l^{\mathrm{S}}\in\{1,2,\cdots,K\}$. 教师网络第$l^{\mathrm{T}}$个辅助分支对学生网络第$l^{\mathrm{S}}$个辅助分支进行图表征蒸馏时,误差项规则化为

$$L_{(l^{\mathrm{T}},l^{\mathrm{S}})}\left(\boldsymbol{H}_{l^{\mathrm{T}}}^{\mathrm{T}},\boldsymbol{H}_{l^{\mathrm{S}}}^{\mathrm{S}}\right)=\alpha L_{\mathrm{MSE}}\left(\boldsymbol{H}_{l^{\mathrm{T}}}^{\mathrm{T}},\boldsymbol{H}_{l^{\mathrm{S}}}^{\mathrm{S}}\right)+\beta L_{\mathrm{ED}}\left(\boldsymbol{H}_{l^{\mathrm{T}}}^{\mathrm{T}},\boldsymbol{H}_{l^{\mathrm{S}}}^{\mathrm{S}}\right)+\gamma L_{\cos}\left(\boldsymbol{H}_{l^{\mathrm{T}}}^{\mathrm{T}},\boldsymbol{H}_{l^{\mathrm{S}}}^{\mathrm{S}}\right) \tag{16}$$

其中,α、β和γ分别是3个误差项的权重系数.

对于每一个学生网络辅助分支产生的图特征向量$\boldsymbol{H}_{l^{\mathrm{S}}}^{\mathrm{S}},l^{\mathrm{S}}\in\{1,2,\cdots,K\}$,本方法使用教师网络的$K$个辅助分支产生的图特征向量$\boldsymbol{H}_{l^{\mathrm{T}}}^{\mathrm{T}}(l^{\mathrm{T}}=1,2,3,\cdots,K)$,通过图表征蒸馏的方式传递知识. 因此,图表征的整体误差项公式为

$$L_{\mathrm{GR}}\left(f^{\mathrm{S}}|X,Y,\eta,f^{\mathrm{T}}\right)=\sum_{l^{\mathrm{S}}=1}^{K}\sum_{l^{\mathrm{T}}=1}^{K}\lambda_{(l^{\mathrm{T}},l^{\mathrm{S}})}L_{(l^{\mathrm{T}},l^{\mathrm{S}})}\left(\boldsymbol{H}_{l^{\mathrm{T}}}^{\mathrm{T}},\boldsymbol{H}_{l^{\mathrm{S}}}^{\mathrm{S}}\right) \tag{17}$$

其中,$X=\{\boldsymbol{x}_i\}(i=1,2,\cdots,N)$代表$N$个输入样本集合,$Y=\{y_i\}(i=1,2,\cdots,N)$代表$N$个输入样本的对应类别标签集合,$\lambda_{(l^{\mathrm{T}},l^{\mathrm{S}})}$代表教师网络的第$l^{\mathrm{T}}$个辅助分支对学生网络的第$l^{\mathrm{S}}$个辅助分支进行图表征蒸馏误差的匹配权重. 本方法通过元网络η在线训练得到$\lambda_{(l^{\mathrm{T}},l^{\mathrm{S}})}$. 该误差项是在给定输入样本信息$X$和$Y$、元网络$\eta$和教师网络$f^{\mathrm{T}}$的条件下对学生网络$f^{\mathrm{S}}$进行图表征蒸馏的误差项. 下一节将讲解怎样采用元学习[28]的方式对图表征误差项和元网络进行交替优化.

3.4.2　利用元学习优化图表征蒸馏

给定输入样本信息X和Y,学生网络利用交叉熵(Cross Entropy,CE)函数来学习输入样本类别标签的映射关系,公式为

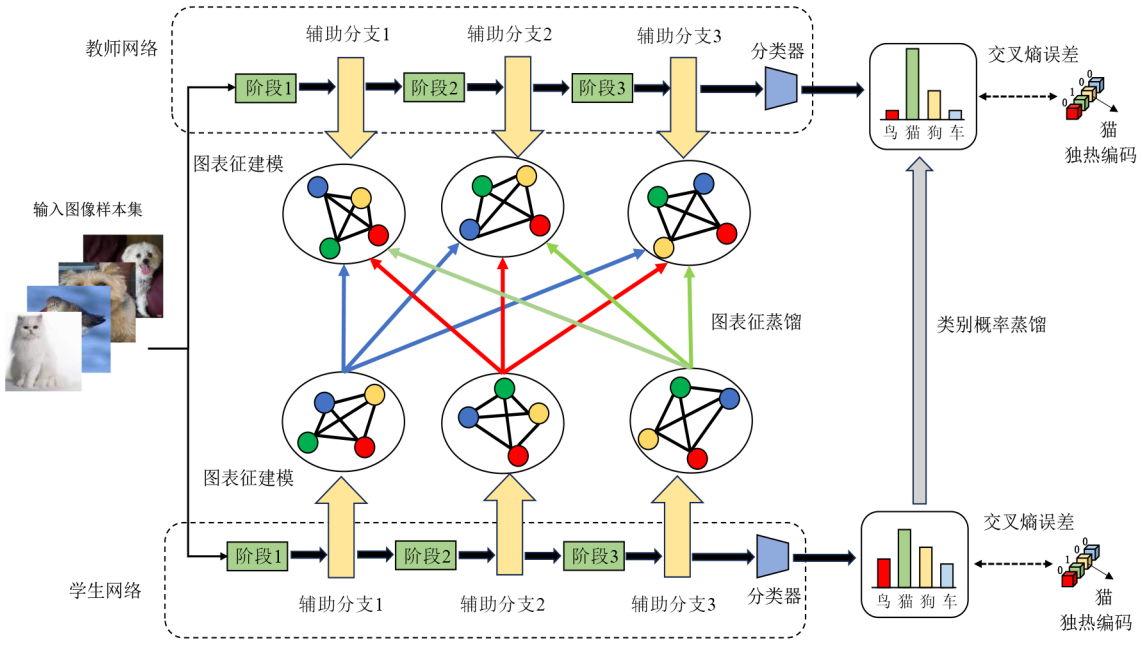


图1　基于图表征知识蒸馏的图像分类方法整体示意图

$$L_{\mathrm{CE}}\left(f^{\mathrm{S}}|X,Y\right)=-\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\tau_{c,y_i}\log p_{i,c} \tag{18}$$

其中，τ_{c,y_i}是一个指示器函数，如果$c=y_i$，则$\tau_{c,y_i}=1$，否则$\tau_{c,y_i}=0$. 交叉熵误差作为任务误差可引导网络学习到相关的分类特征信息，为图表征蒸馏提供基础. 本文将式(18)的交叉熵任务误差与式(17)的图表征蒸馏误差作为特征层面的蒸馏误差，表达为

$$L_{\mathrm{total}}\left(f^{\mathrm{S}}|X,Y,\eta,f^{\mathrm{T}}\right)=L_{\mathrm{CE}}\left(f^{\mathrm{S}}|X,Y\right)+L_{\mathrm{GR}}\left(f^{\mathrm{S}}|X,Y,\eta,f^{\mathrm{T}}\right) \tag{19}$$

本方法的目标是训练学生网络f^{S}获得优越的分类表现. 为实现这个目标，关键的图表征蒸馏误差$L_{\mathrm{GR}}\left(f^{\mathrm{S}}|X,Y,\eta,f^{\mathrm{T}}\right)$需要使学生网络从教师网络中学习到更好的特征表达. 为匹配教师-学生网络辅助分支之间的蒸馏，本方法引入元网络η来与图表征蒸馏误差进行交替优化，具体过程如下：

(1)H次最小化$L_{\mathrm{GR}}\left(f^{\mathrm{S}}|X,Y,\eta,f^{\mathrm{T}}\right)$来训练$f^{\mathrm{S}}$；

(2)单次最小化$L_{\mathrm{CE}}\left(f^{\mathrm{S}}|X,Y\right)$来训练$f^{\mathrm{S}}$；

(3)衡量$L_{\mathrm{CE}}\left(f^{\mathrm{S}}|X,Y\right)$并且更新元网络$\eta$.

在第一阶段，本方法通过H次最小化$L_{\mathrm{GR}}\left(f^{\mathrm{S}}|X,Y,\eta,f^{\mathrm{T}}\right)$来优化初始学生网络$f^{\mathrm{S}}$的参数$\theta_{\mathrm{S}}^{0}$，得到的$\theta_{\mathrm{S}}^{H}$是从特征层面的图表征蒸馏中优化而来，因此，更凸显了蒸馏层匹配权重$\lambda_{(l^{\mathrm{T}},l^{\mathrm{S}})}$的重要. 本方法取$H=2$就可以达到元优化目标. 第二阶段是通过最小化交叉熵任务误差$L_{\mathrm{CE}}\left(f^{\mathrm{S}}|X,Y\right)$完成从$\theta_{\mathrm{S}}^{H}$到$\theta_{\mathrm{S}}^{H+1}$的单步适应. 第三阶段，任务导向的误差$L_{\mathrm{CE}}\left(\theta_{\mathrm{S}}^{H+1}|X,Y\right)$通过第一阶段和第二阶段用到的输入样本评估了网络从$\theta_{\mathrm{S}}^{H}$到$\theta_{\mathrm{S}}^{H+1}$的适应程度. 最后，元网络$\eta$通过最小化误差$L_{\mathrm{CE}}\left(\theta_{\mathrm{S}}^{H+1}|X,Y\right)$进行优化. 综上所述，元网络的优化目标公式为

$$\begin{aligned}&\min\nolimits_{\eta}L_{\mathrm{CE}}\left(\theta_{\mathrm{S}}^{H+1}|X,Y\right)\\ &\text{s.t. }\theta_{\mathrm{S}}^{H+1}=\theta_{\mathrm{S}}^{H}-\nabla_{\theta_{\mathrm{S}}}L_{\mathrm{CE}}\left(\theta_{\mathrm{S}}^{H}|X,Y\right)\\ &\theta_{\mathrm{S}}^{h+1}=\theta_{\mathrm{S}}^{h}-\nabla_{\theta_{\mathrm{S}}}L_{\mathrm{GR}}\left(\theta_{\mathrm{S}}^{h}|X,Y,\eta,f^{\mathrm{T}}\right)\\ &h=0,1,\cdots,H-1\end{aligned} \tag{20}$$

元优化采用交替更新学生网络f^{S}和元网络η的方式进行. 首先，通过最小化$L_{\mathrm{total}}\left(f^{\mathrm{S}}|X,Y,\eta,f^{\mathrm{T}}\right)$训练学生网络$f^{\mathrm{S}}$的参数$\theta_{\mathrm{S}}$，然后，通过3个阶段优化来训练元网络$\eta$.

3.4.3 元网络的结构设计

元网络$\eta=\left\{\left\{\boldsymbol{\eta}_{l^{\mathrm{T}}}\in\mathbb{R}^{D\times D}\right\}\cup\left\{\boldsymbol{\eta}_{l^{\mathrm{S}}}\in\mathbb{R}^{D\times D}\right\}\right\}$包含$K$个教师和$K$个学生线性转换矩阵. 对于匹配的第$l^{\mathrm{T}}$个教师网络的辅助分支与第$l^{\mathrm{S}}$个学生网络的辅助分支，$l^{\mathrm{T}}\in\{1,2,\cdots,K\}$，$l^{\mathrm{S}}\in\{1,2,\cdots,K\}$，输出的图特征分别是$\boldsymbol{H}_{l^{\mathrm{T}}}^{\mathrm{T}}\in\mathbb{R}^{N\times D}$和$\boldsymbol{H}_{l^{\mathrm{S}}}^{\mathrm{S}}\in\mathbb{R}^{N\times D}$. 元网络分别采用$\boldsymbol{\eta}_{l^{\mathrm{T}}}$和$\boldsymbol{\eta}_{l^{\mathrm{S}}}$来对教师和学生图特征进行转换，并使用$l_2$规范化分别得到教师特征$\boldsymbol{S}_{l^{\mathrm{T}}}^{\mathrm{T}}$和学生特征$\boldsymbol{S}_{l^{\mathrm{S}}}^{\mathrm{S}}$：

$$\boldsymbol{S}_{l^{\mathrm{T}}}^{\mathrm{T}}=\boldsymbol{H}_{l^{\mathrm{T}}}^{\mathrm{T}}\boldsymbol{\eta}_{l^{\mathrm{T}}}\in\mathbb{R}^{N\times D},\ \boldsymbol{S}_{l^{\mathrm{S}}}^{\mathrm{S}}=\boldsymbol{H}_{l^{\mathrm{S}}}^{\mathrm{S}}\boldsymbol{\eta}_{l^{\mathrm{S}}}\in\mathbb{R}^{N\times D}$$

使用$\boldsymbol{S}_{l^{\mathrm{T}}}^{\mathrm{T}}$和$\boldsymbol{S}_{l^{\mathrm{S}}}^{\mathrm{S}}$之间的余弦相似度作为蒸馏层匹配的依据，并使用S型曲线(Sigmoid)激活函数将余弦相似度归一化到(0,1)的范围内，公式表达如下：

$$\lambda_{(l^{\mathrm{T}},l^{\mathrm{S}})}=\frac{1}{1+\exp\left(-\dfrac{\boldsymbol{S}_{l^{\mathrm{T}}}^{\mathrm{T}}\cdot\boldsymbol{S}_{l^{\mathrm{S}}}^{\mathrm{S}}}{\left\|\boldsymbol{S}_{l^{\mathrm{T}}}^{\mathrm{T}}\right\|\left\|\boldsymbol{S}_{l^{\mathrm{S}}}^{\mathrm{S}}\right\|}\right)} \tag{21}$$

3.4.4 学生网络整体误差构建

除了图表征蒸馏误差，本文在教师-学生网络的辅助分支之间采用类别概率蒸馏误差，引导学生网络学习到教师网络的类别决策信息. 给定输入样本集$X=\left\{\boldsymbol{x}_i\right\}(i=1,2,3,\cdots,N)$，教师-学生网络$K$个辅助分支产生的类别概率分布矩阵分别为$\boldsymbol{p}_l^{\mathrm{T}}\in\mathbb{R}^{N\times C}(l=1,2,3,\cdots,K)$和$\boldsymbol{p}_l^{\mathrm{S}}\in\mathbb{R}^{N\times C}(l=1,2,3,\cdots,K)$，其中，$N$个行向量代表$N$个样本的类别概率分布. 本方法使用相对熵引导学生网络K个辅助分支产生的类别概率分布去学习教师网络的类别概率分布，公式为

$$L_{\mathrm{KL}}\left(f^{\mathrm{S}}|X,Y,f^{\mathrm{T}}\right)=\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C}\boldsymbol{p}_l^{\mathrm{T}}[n][c]\log\frac{\boldsymbol{p}_l^{\mathrm{T}}[n][c]}{\boldsymbol{p}_l^{\mathrm{S}}[n][c]} \tag{22}$$

其中，n和c分别代表矩阵的行索引和列索引.

图1展示了训练学生网络的整体示意图，包括基础的交叉熵任务误差、图表征蒸馏误差以及类别概率蒸馏误差. 最终的总误差表达为

$$\begin{aligned}L\left(f^{\mathrm{S}}|X,Y,\eta,f^{\mathrm{T}}\right)=&L_{\mathrm{CE}}\left(f^{\mathrm{S}}|X,Y\right)+L_{\mathrm{GR}}\left(f^{\mathrm{S}}|X,Y,\eta,f^{\mathrm{T}}\right)\\&+L_{\mathrm{KL}}\left(f^{\mathrm{S}}|X,Y,f^{\mathrm{T}}\right)\end{aligned} \tag{23}$$

整体的算法伪代码如算法1所示：

算法1 基于图表征知识蒸馏的图像分类方法

输入：教师网络f^{T}、输入样本集X和Y

输出：学生网络f^{S}和元网络η

初始化$\theta_{\mathrm{S}}^{H}\leftarrow\theta_{\mathrm{S}}$

FOR $h=0,1,\cdots,H-1$:

更新学生网络f^{S}:$\theta_{\mathrm{S}}^{h+1}\leftarrow\theta_{\mathrm{S}}^{h}-\nabla_{\theta_{\mathrm{S}}}L_{\mathrm{GR}}\left(\theta_{\mathrm{S}}^{h}|X,Y,\eta,f^{\mathrm{T}}\right)$

END FOR

更新学生网络f^{S}:$\theta_{\mathrm{S}}^{H+1}\leftarrow\theta_{\mathrm{S}}^{H}-\nabla_{\theta_{\mathrm{S}}}L_{\mathrm{CE}}\left(\theta_{\mathrm{S}}^{H}|X,Y\right)$

更新元网络η:$\eta\leftarrow\eta-\nabla_{\eta}L_{\mathrm{CE}}\left(\theta_{\mathrm{S}}^{H+1}|X,Y\right)$

更新学生网络f^{S}:$\theta_{\mathrm{S}}\leftarrow\theta_{\mathrm{S}}-\nabla_{\theta_{\mathrm{S}}}L\left(f^{\mathrm{S}}|X,Y,\eta,f^{\mathrm{T}}\right)$

4 实验

4.1 图像分类数据集

本文在实验中使用了3种图像分类数据集，分别是CIFAR-100、STL-10和ImageNet. 数据集的具体介绍

如下:

CIFAR-100数据集:该数据集是加拿大高等研究院发布的100种分类自然图像数据集,主要包含动物、植物、家用工具和交通工具等目标类别. 该数据集的训练集包含5×104张图像,测试集包含1×104张图像.

STL-10数据集:该数据集是用于自主学习(Self-Taught Learning,STL)的10种分类自然图像数据集,主要包含了猫、狗和鸟等动物类别以及飞机、轮船和车辆等交通工具类别. 该数据集的训练集包含5 000张图像,测试集包含8 000张图像.

ImageNet数据集:该数据集是1个大规模的图像分类数据集,包含了1 000个目标类别,包括各种不同的动物、植物、物体和场景. 该数据集的训练集包含128×10^4张图像,测试集包含5×10^4张图像.

4.2　实验设置

4.2.1　CIFAR-100数据集上的实验设置

本文采用随机裁剪和翻转来对输入图像进行数据增强,将图像的输入尺寸标准化为32×32,采用随机梯度下降优化器对网络进行训练,动量系数为0.9,批数据规模为64,权重衰减因子为5×10^{-4}. 总的训练轮数为240,学习率初始化为0.05,到第150轮和180轮进行1/10衰减.

4.2.2　STL-10数据集上的实验设置

本文采用随机裁剪和翻转来对输入图像进行数据增强,将图像的输入尺寸标准化为32×32,采用随机梯度下降优化器对网络进行训练,动量系数为0.9,批数据规模为64,权重衰减因子为0. 总共的训练轮数为30,学习率初始化为0.1,到第30轮、第60轮和第90轮时进行1/10衰减.

4.2.3　ImageNet数据集上的实验设置

本文采用随机裁剪和翻转来对输入图像进行数据增强,将图像的输入尺寸标准化为224×224,采用随机梯度下降优化器对网络进行训练,动量系数为0.9,批数据规模为256,权重衰减因子为1×10^{-4}. 总的训练轮数为100,学习率初始化为0.1,到第30轮、第60轮和第90轮时进行1/10衰减.

4.2.4　评价指标

本文采用准确率来衡量图像分类的性能,准确率的定义是测试集中分类正确的样本数量与总样本数量的比例. 准确率越高,表明分类性能越好.

4.2.5　图像分类网络结构

本文在广泛的网络结构上进行蒸馏实验来验证本文提出方法的有效性,包括残差网络(Residual Networks,ResNet)[2]、广度残差网络(Wide Residual Networks,WRN)[29]、牛津大学视觉几何组网络(Visual Geometry Group,VGG)[30]、第二版本移动端网络(Mobile Networks Version 2,MobileNetV2)[31]和乱序网络(Shuffle Networks,ShuffleNet)[32]. 采用的教师-学生网络结构组合以及参数量信息如表1所示. 其中,本文用度量单位Million(M)来衡量参数的量级. 从表1可以看出,教师网络相比学生网络具有更多参数量,结构更复杂.

表1　教师网络&学生网络组合的参数量对比

教师网络&学生网络	参数量/M
WRN-40-2 & WRN-16-2	2.26 & 0.70
WRN-40-2 & WRN-40-1	2.26 & 0.57
ResNet-56 & ResNet-20	0.86 & 0.28
ResNet-32x4 & ResNet-8x4	7.43 & 1.23
VGG-13 & MobileNetV2	9.46 & 0.81
ResNet-50 & MobileNetV2	23.71 & 0.81
WRN-40-2 & ShuffleNetV1	2.26 & 0.95
ResNet-32x4 & ShuffleNetV2	7.43 & 1.36
ResNet-34 & ResNet-18	21.80 & 11.69

4.2.6　对比的知识蒸馏方法

本文与现有的多种知识蒸馏方法进行对比,包括传统的知识蒸馏(Knowledge Distillation,KD)[6]、拟合网络(Fit Network,FitNet)[7]、注意力转移(Attention Transfer,AT)[8]、变分信息蒸馏(Variational Information Distillation,VID)[9]、关系知识蒸馏(Relational Knowledge Distillation,RKD)[11]、对比表征蒸馏(Contrastive Representation Distillation,CRD)[10]和解耦知识蒸馏(Decoupled Knowledge Distillation,DKD)[14].

4.3　实验结果

4.3.1　CIFAR-100数据集上图像分类结果

表2展示了在不同的教师-学生网络结构组合上应用各种知识蒸馏方法在CIFAR-100数据集上的图像分类准确率对比. 表2~4中的加粗数字代表最高的准确率. 从表中可以看出,本文提出的GRKD在所有教师-学生网络组合上获得了一致最佳性能,表明提出的GRKD可以通过图表征的方式有效引导学生网络学习到教师的知识.

相比于没有知识蒸馏的基线,本文提出的GRKD在表2中的8个教师-学生网络组合上获得了3.70%的性能提升. 此外,本文提出的GRKD凭借平均准确率1.60%的提升超越了先前最好的DKD方法. 即使是先前最好的DKD方法,也并没有在8个网络结构上稳定超越之前的对比方法. 比如,当教师网络为WRN-40-2,学生网络为ShuffleNetV1时,CRD方法训练的学生网络准确率为72.69%,高于DKD方法的72.46%,低于GRKD的74.08%. 作为对比,本文提出的GRKD方法在提升所有学生网络的准确率上有更好鲁棒性.

表2 在不同的教师-学生结构组合上各种知识蒸馏方法在CIFAR-100数据集上图像分类准确率对比 单位:%

教师网络结构		学生网络	学生网络准确率								
名称	准确率	结构名称	基线	KD	FitNet	AT	VID	RKD	CRD	DKD	GRKD
WRN-40-2	76.73	WRN-16-2	72.58	73.64	74.18	74.43	74.27	74.66	74.78	74.89	**76.21**
WRN-40-2	76.73	WRN-40-1	71.01	72.26	72.44	72.73	72.79	72.75	73.14	73.36	**74.95**
ResNet-56	73.64	ResNet-20	68.88	70.22	70.84	71.11	70.68	71.09	71.15	71.19	**72.68**
ResNet-32x4	79.45	ResNet-8x4	72.52	73.07	74.83	74.69	74.21	74.42	75.23	75.29	**77.24**
VGG-13	74.79	MobileNetV2	68.67	69.62	70.06	69.83	70.26	69.94	70.34	70.12	**71.71**
ResNet-50	79.09	MobileNetV2	68.67	69.63	70.24	70.35	70.34	70.49	70.84	70.68	**72.15**
WRN-40-2	76.73	ShuffleNetV1	70.52	71.25	71.38	71.75	71.83	72.14	72.69	72.46	**74.08**
ResNet-32x4	79.45	ShuffleNetV2	70.56	71.18	71.42	71.37	71.95	71.79	71.86	72.25	**73.99**

4.3.2 迁移到STL-10数据集上图像分类结果

GRKD蒸馏得到的学生网络在CIFAR-100数据集上获得了最佳表现. 除了在目标数据集上获得准确识别率,本文进一步评测在CIFAR-100数据集上蒸馏训练的学生网络对于未见过的STL-10图像分类数据集的性能,以此评估蒸馏的泛化能力. 本文使用特征线性评测协议,固定在上游CIFAR-100数据集上训练的特征提取器部分,只训练1个额外的线性分类器对下游的STL-10数据集进行图像分类任务.

表3展示了不同学生结构上各种知识蒸馏方法迁移到下游STL-10数据集上的图像分类准确率. 从表中可看出,本文提出的GRKD在所有教师-学生网络组合上获得最佳性能. 相比于基线,本文提出的GRKD在表3中的8个教师-学生网络组合上获得了3.76%的性能提升. 相比于先前最好的DKD方法,本文提出的GRKD获得了2.20%的性能提升. 实验结果表明提出的GRKD可通过图表征的方式引导学生网络学习到更加泛化的视觉表征,使学习到的特征即使迁移到其他未知的图像分类数据集也能获得很好性能.

表3 在不同的学生结构上各种知识蒸馏方法迁移到下游STL-10数据集上图像分类准确率对比 单位:%

学生网络	学生网络准确率								
结构名称	基线	KD	FitNet	AT	VID	RKD	CRD	DKD	GRKD
WRN-16-2	66.88	67.12	67.34	67.36	67.63	67.24	67.94	67.69	**69.75**
WRN-40-1	62.50	63.42	63.63	63.85	63.72	63.93	64.06	64.11	**66.33**
ResNet-20	62.96	63.75	63.88	64.21	64.36	64.52	64.46	64.22	**66.49**
ResNet-8x4	66.89	68.27	68.33	68.53	68.62	69.24	69.53	69.83	**72.36**
MobileNetV2	67.74	68.31	68.53	68.68	68.53	68.87	68.83	68.75	**71.00**
MobileNetV2	67.74	68.65	68.75	68.97	69.23	69.14	69.52	69.72	**71.90**
ShuffleNetV1	67.79	68.23	68.65	68.73	68.83	68.77	69.24	69.31	**71.37**
ShuffleNetV2	68.93	69.52	69.67	69.83	70.13	70.31	70.43	70.28	**72.30**

4.3.3 ImageNet数据集上图像分类结果

将提出的GRKD方法用于大规模的ImageNet图像分类实验. 表4展示ResNet-50为教师网络,ResNet-18为学生网络的情况下,不同知识蒸馏方法在ImageNet数据集上的图像分类准确率结果对比. 如表4所示,本文提出的GRKD获得了最佳性能,在ResNet-18网络准确率达到了71.24%,分别凭借1.40%和0.57%的准确率提升超越了基线和先前最好的DKD方法. 实验结果显示了GRKD方法不仅在小规模的CIFAR-100图像分类任务中取得很好效果,在大规模1 000类别图像分类任务中依旧能够获得优越性能,表明图表征蒸馏方法可以有效提升学生网络对分类任务的泛化能力.

表4 知识蒸馏方法在ImageNet数据集上图像分类准确率对比 单位:%

教师网络结构		学生网络	学生网络准确率								
名称	准确率	结构名称	基线	KD	FitNet	AT	VID	RKD	CRD	DKD	GRKD
ResNet-50	76.13	ResNet-18	69.84	70.31	70.46	70.52	70.48	70.63	70.55	70.67	**71.24**

4.3.4 数据缺乏场景下的图像分类结果

在实践中,训练样本的数量是不充足的,因此,本章节评估了知识蒸馏方法在样本数量缺乏条件下的性能. 图2展示了不同知识蒸馏方法在25%、50%、75%和100%训练样本条件下的CIFAR-100图像分类结果,其中教师网络为ResNet-32x4,学生网络为ResNet-8x4,准

确率基于原始的CIFAR-100测试集. 本实验采用分层采样的策略对训练样本进行划分来保证训练子集是类别平衡的. 对每一种小样本场景,不同方法都采用的是相同训练数据确保公平性.

从图2可以看出,本文提出的GRKD在不同比例的训练样本条件下超越了其他知识蒸馏方法. 相比于没有知识蒸馏的基线,提出的GRKD在保留25%、50%、75%和100%训练样本条件下分别获得了3.89%、3.20%、2.35%和4.72%的性能提升. 相比于先前最好的DKD方法,GRKD分别获得了1.83%、1.92%、1.49%和1.95%的准确率提升. 实验结果显示了GRKD即使在数据缺乏的条件下依旧获得很好的分类性能. 这表明图表征蒸馏方法可以通过结点间建模方式从有限数据中学习到通用的特征表达. 先前的知识蒸馏方法通常将单一样本作为知识来源,导致学生网络在数据缺乏场景下容易陷入过拟合,在测试集上表现较差.

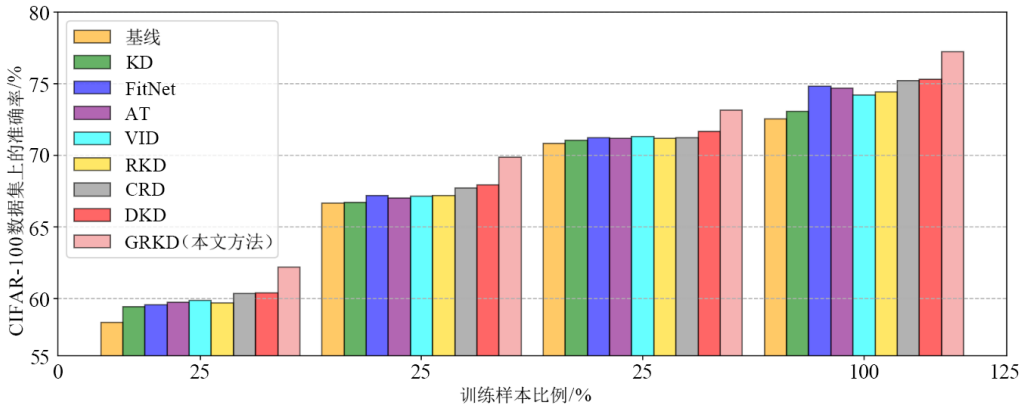


图2　知识蒸馏方法在CIFAR-100数据集不同比例训练样本上的图像分类准确率对比

4.4　消融实验与参数分析

4.4.1　蒸馏误差消融实验

本文提出的蒸馏方法包括3个误差项,分别是:交叉熵任务误差L_{CE}、图表征蒸馏误差L_{GR}和类别概率蒸馏误差L_{KL}. 表5展示了对误差项的消融实验,其中,实验的教师网络为ResNet-32x4,学生网络为ResNet-32x8.

表5　CIFAR-100数据集上的误差消融实验　　单位:%

误差项	准确率
交叉熵任务误差L_{CE}(基线)	72.52
图表征蒸馏误差L_{GR}	76.87
类别概率蒸馏误差L_{KL}	74.73
整体误差L_{CE}+L_{GR}+L_{KL}	**77.24**

基础的交叉熵任务误差L_{CE}训练可以使学生网络表现出72.52%的分类准确率,在此基础上分别应用图表征蒸馏误差L_{GR}和类别概率蒸馏误差L_{KL}使学生网络的准确率分别提升到76.87%和74.73%. 实验结果证明特征层面的图表征蒸馏效果显著好于概率层面的类别概率蒸馏. 最后,将2种蒸馏误差相加作为整体误差可最大程度提升学生网络,准确率达到77.24%,比没有蒸馏的基线提升了4.72%.

4.4.2　基于元优化的层匹配机制消融实验

本文采用全连接的层匹配机制,并基于元学习来优化层匹配权重. 本章节对比了传统一对一和全连接匹配机制,实验结果如表6所示,其中,本实验的教师网络为ResNet-32x4,学生网络为ResNet-32x8. 一对一指的是教师-学生网络基于相同层特征进行匹配,且匹配权重为1. 传统的全连接匹配模式采用匹配权重为1,并没有采用加权形式. 从表6中可以看出,传统的一对一匹配机制要好于全连接模式,因为,传统的全连接匹配模式没有对权重进行自适应优化,导致层匹配会出现语义不一致问题,导致图表征建模效果下降.

本文提出的元优化匹配权重相比传统一对一和全连接匹配分别获得了1.74%和2.02%的准确率提升. 这是因为元学习可以根据教师-学生网络的特征相似度自适应匹配建模,完成自适应的图蒸馏层匹配.

表6　CIFAR-100数据集上的层匹配机制消融实验　　单位:%

层匹配机制	准确率
一对一	75.13
全连接	74.85
基于元优化的加权全连接	**76.87**

为进一步验证基于元学习的层匹配机制,本章节统计了层匹配权重在初始状态和收敛状态下的分布变化,如图3所示,教师1~3和学生1~3分别代表辅助分支索引. 从初始状态到收敛状态,基于元学习的机制对相同层分配更大的匹配权重,对不同层分配适量的权重. 相比于传统的匹配机制,元学习的方式可以通过数据驱动方式自适应捕捉到匹配层之间相关性,使图表征

蒸馏方法可以轻易扩展到任意结构组合的教师-学生网络.

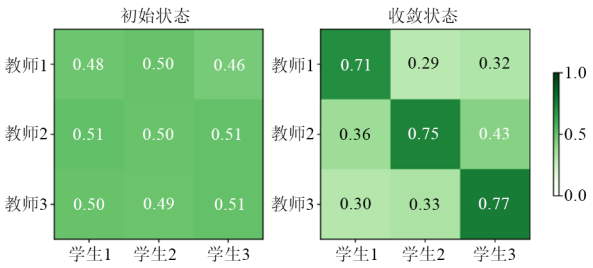


图3 基于元优化的图表征蒸馏层匹配权重

4.4.3 图蒸馏误差权重超参数分析

本文提出的图蒸馏误差包括:基于最小平方误差的图特征蒸馏L_{MSE}、基于欧氏距离的图关系蒸馏L_{ED}、基于三元组夹角的图关系蒸馏L_{cos}3个误差项.本章节对3种误差项的权重超参数进行实验,教师网络为ResNet-32x4,学生网络为ResNet-32x8.如图4所示,图4(a)~(c)分别为α、β和γ在不同权重值下的准确率曲线.实验结果显示当权重值的范围在1~10之间时,3种误差的分类准确率会逐渐上升并在权重等于10时到达最高值.当权重值大于10的时候,随着权重增加,3种误差的分类准确率会逐渐下降,表明误差出现了过优化现象导致性能衰退.因此,本文取$\alpha=10$、$\beta=10$和$\gamma=10$作为误差权重超参数.

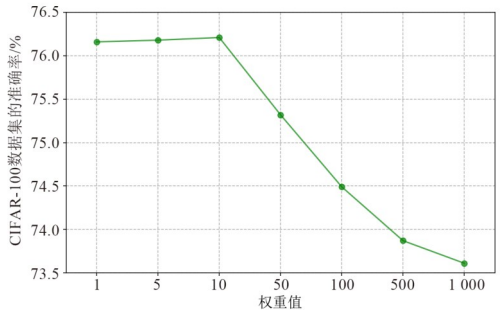
4.4.4 图表征蒸馏误差消融实验

本章节对图表征蒸馏的3个误差项进行消融实验.教师网络为ResNet-32x4,学生网络为ResNet-32x8.

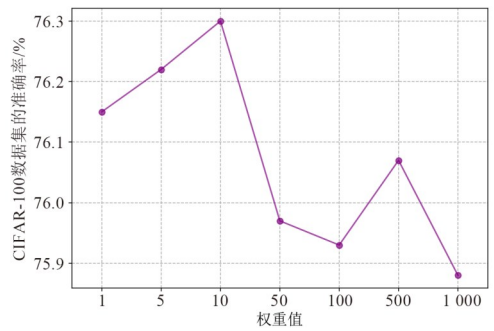
从表7可以看出,在基线之上分别使用基于最小平方误差的图特征蒸馏L_{MSE}、基于欧氏距离的图关系蒸馏L_{ED}和基于三元组夹角的图关系蒸馏L_{cos},可以获得3.69%、3.78%和3.86%的准确率提升,证明每一个蒸馏误差项都使学生网络得到显著性能增益.图关系蒸馏的误差组合$L_{cos}+L_{ED}$相比图特征蒸馏误差进一步提升了0.51%的准确率,图关系相比图特征蒸馏对网络的分类性能提升更加显著.最终,将图特征和图关系蒸馏联合作为整体误差使学生网络获得最高76.87%的准确率.

4.4.5 特征空间可视化

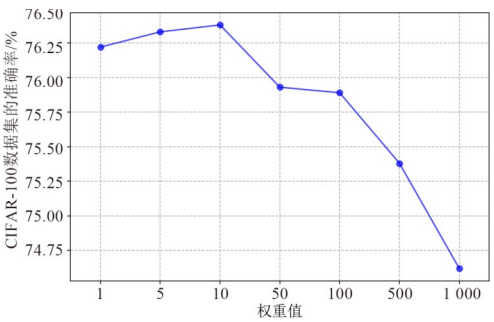
如图5(a)和图5(b)所示,本章节分别对先前最好的DKD蒸馏方法和本文提出的GRKD方法优化得到的特征进行t分布式随机邻居嵌入(t-distributed Stochastic Neighbor Embedding,t-SNE)[33]可视化.图5由10种颜色构成,分别代表CIFAR-10数据集中的10个类别.从图5中可以看出,相比于DKD方法,本文提出的GRKD方法训练得到的学生网络产生的特征空间具有更好的类内紧实度和类间分离度,表明图表征蒸馏可产生更好的判别性特征,有利于下游图像分类.



(a) 误差L_{MSE}的权重α



(b) 误差L_{ED}的权重β



(c) 误差L_{cos}的权重γ

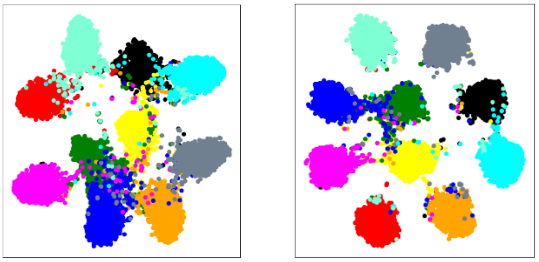
图4 CIFAR-100数据集上的误差项权重敏感性实验

表7 CIFAR-100数据集上图表征蒸馏误差消融实验 单位:%

误差项	准确率
交叉熵任务误差L_{CE}(基线)	72.52
最小平方误差的图特征蒸馏L_{MSE}	76.21
基于欧氏距离的图关系蒸馏L_{ED}	76.30
基于三元组夹角的图关系蒸馏L_{cos}	76.38
误差组合$L_{cos}+L_{ED}$	76.72
整体误差$L_{CE}+L_{GR}+L_{KL}$	**76.87**

4.4.6 图像热力图可视化

如图6所示,DKD方法和本文提出的GRKD方法对输入图像的热力图进行可视化,采用梯度类别激活图[34](Gradient-Class Activation Map,Grad-CAM),红色



(*a*) DKD方法的特征空间　　(*b*) GRKD方法的特征空间

图5　知识蒸馏方法对学生网络的特征空间可视化效果对比

区域代表模型重点关注部分,蓝色区域代表模型不关注部分. 从图6可以看出,GRKD方法相比DKD更能捕捉到具有判别性的物体区域和特征,获得更好的图像分类能力.

例如,对于输入的牧羊犬图像,DKD方法将犬的嘴部作为分类依据,而GRKD则将全局的犬脸作为更具判别性的区域. 对于输入的水库图像,GRKD方法也强调了更加全局的信息区域. 对输入的朱雀图片,其背景中有较多树林干扰,对DKD的识别有较大负面影响,但GRKD却能够从噪声中准确定位到朱雀目标. 对于菌菇类别,DKD方法没有正确定位到目标,而GRKD则正确捕捉到了目标区域,正确分类.

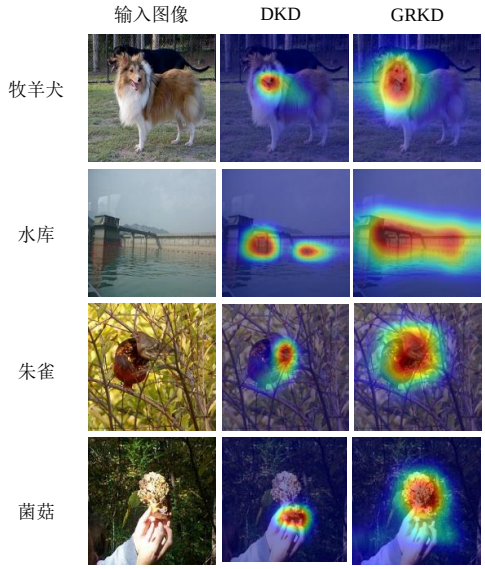


图6　输入图像的热力图可视化效果对比

5　结束语

本文提出了一种基于图表征知识蒸馏方法用于图像分类任务. 该方法通过引入图卷积神经网络来建模样本间的关系,捕捉到全局的样本依赖,提升对样本的特征学习能力. 图表征知识蒸馏引导学生网络从图结点特征和图关系角度更好地学习教师网络的图信息,提升学生网络的图建模能力. 在CIFAR-100、STL-10以及ImageNet数据集上的实验结果表明本文提出的GRKD显著超越了先前的知识蒸馏方法,证明GRKD可以通过图表征蒸馏的方式引导学生网络学习具有判别性的特征空间,提升图像分类效果.

本文提出了基于图表征的知识蒸馏方法,采用基于相似度的图建模框架. 希望未来的知识蒸馏工作能提出更好的图结点建模方法提升图像分类任务.

参考文献

[1] 王波, 黄冕, 刘利军, 等. 基于多层聚焦Inception-V3卷积网络的细粒度图像分类[J]. 电子学报, 2022, 50(1): 72-78.
WANG B, HUANG M, LIU L J, et al. Multi-layer focused Inception-V3 models for fine-grained visual recognition[J]. Acta Electronica Sinica, 2022, 50(1): 72-78. (in Chinese)

[2] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Computer Society, 2016: 770-778.

[3] ZHU H, AN Z L, YANG C G, et al. EENA: Efficient evolution of neural architecture[C]//2019 International Conference on Computer Vision Workshop (ICCVW). Piscataway: IEEE, 2019: 1891-1899.

[4] YANG C G, AN Z L, LI C, et al. Multi-objective pruning for CNNs using genetic algorithm[C]//International Conference on Artificial Neural Networks. Cham: Springer, 2019: 299-305.

[5] 郑云飞, 王晓兵, 张雄伟, 等. 基于金字塔知识的自蒸馏HRNet目标分割方法[J]. 电子学报, 2023, 51(3): 746-756.
ZHENG Y F, WANG X B, ZHANG X W, et al. The self-distillation HRNet object segmentation based on the pyramid knowledge[J]. Acta Electronica Sinica, 2023, 51(3): 746-756. (in Chinese)

[6] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[EB/OL]. (2015-03-05)[2023-09-02]. https://arxiv.org/abs/1503.02531.

[7] ROMERO A, BALLAS N, KAHOU S E, et al. Fitnets: Hints for thin deep nets[EB/OL]. (2014-12-19)[2023-09-05]. http://arxiv.org/abs/1412.6550.

[8] ZAGORUYKO S, KOMODAKIS N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer[EB/OL]. (2016-12-12)[2023-09-05]. https://arxiv.org/abs/1612.03928.

[9] AHN S, HU S X, DAMIANOU A, et al. Variational information distillation for knowledge transfer[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recogni-

tion (CVPR). Piscataway: IEEE, 2019: 9155-9163.

[10] TIAN Y L, KRISHNAN D, ISOLA P. Contrastive representation distillation[EB/OL]. (2019-10-23) [2023-9-16]. http://arxiv.org/abs/1910.10699.

[11] PARK W, KIM D, LU Y, et al. Relational knowledge distillation[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 3962-3971.

[12] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[EB/OL]. (2016-09-23) [2023-09-06]. https://arxiv.org/abs/1609.02907.

[13] KRIZHEVSKY A. Learning multiple layers of features from tiny images[EB/OL]. (2009-04-08)[2023-10-08]. http://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf.

[14] ZHAO B R, CUI Q, SONG R J, et al. Decoupled knowledge distillation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 11943-11952.

[15] COATES A, NG A Y, LEE H. An analysis of single-layer networks in unsupervised feature learning[C]//International Conference on Artificial Intelligence and Statistics. Fort Lauderdale: JMLR.org, 2011: 215-223.

[16] DENG J, DONG W, SOCHER R, et al. ImageNet: A large-scale hierarchical image database[C]//2009 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Computer Society, 2009: 248-255.

[17] FAN W Q, MA Y, LI Q, et al. Graph neural networks for social recommendation[C]// The World Wide Web Conference. New York: ACM, 2019: 417-426.

[18] THEKUMPARAMPIL K K, WANG C, OH S,et al. Attention-based graph neural network for semi-supervised learning[EB/OL]. (2018-03-10)[2023-10-08]. http://arxiv.org/abs/1803.03735.

[19] 车向北, 康文倩, 邓彬, 等. 一种基于图神经网络的SDN路由性能预测模型[J]. 电子学报, 2021, 49(3): 484-491.
CHE X B, KANG W Q, DENG B, et al. A prediction model of SDN routing performance based on graph neural network[J]. Acta Electronica Sinica, 2021, 49(3): 484-491. (in Chinese)

[20] WANG X L, GUPTA A. Videos as space-time region graphs[C]//European Conference on Computer Vision (ECCV). Cham: Springer, 2018: 413-431.

[21] HU H Z, JI D Y, GAN W H, et al. Class-wise dynamic graph convolution for semantic segmentation[EB/OL]. (2020-07-19)[2023-10-08]. http://arxiv.org/abs/2007.09690.

[22] 袁冠, 邴睿, 刘肖, 等. 基于时空图神经网络的手势识别[J]. 电子学报, 2022, 50(4): 921-931.
YUAN G, BING R, LIU X, et al. Spatial-temporal graph neural network based hand gesture recognition[J]. Acta Electronica Sinica, 2022, 50(4): 921-931. (in Chinese)

[23] YANG C G, AN Z L, CAI L H, et al. Hierarchical self-supervised augmented knowledge distillation[C]//Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2021: 1217-1223.

[24] YANG C G, AN Z L, CAI L H, et al. Mutual contrastive learning for visual representation learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(3): 3045-3053.

[25] YANG C G, ZHOU H L, AN Z L, et al. Cross-image relational knowledge distillation for semantic segmentation [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 12309-12318.

[26] YANG T, ZHU S J, CHEN C, et al. Mutualnet: Adaptive convnet via mutual learning from network width and resolution[EB/OL]. (2019-09-27)[2023-10-10]. http://arxiv.org/abs/1909.12978.

[27] AGARAP A F. Deep learning using rectified linear units (relu)[EB/OL]. (2018-03-22)[2023-10-10]. http://arxiv.org/abs/1803.08375.

[28] FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks[C]//34th International Conference on Machine Learning. New York: ACM, 2017: 1126-1135.

[29] ZAGORUYKO S, KOMODAKIS N. Wide residual networks[C]//British Machine Vision Conference 2016. York: BMVA Press, 2016: 1-15.

[30] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04)[2023-10-12]. https://arxiv.org/abs/1409.1556.

[31] SANDLER M, HOWARD A, ZHU M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Computer Society, 2018: 4510-4520.

[32] ZHANG X Y, ZHOU X Y, LIN M X, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE

Computer Society, 2018: 6848-6856.
[33] VAN DER MAATEN L, HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(11): 2579-2605.
[34] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization[J]. International Journal of Computer Vision, 2020, 128(2): 336-359.

作者简介

杨传广　男,1996年4月出生,山东潍坊人.中国科学院计算技术研究所特别研究助理.主要研究方向为人工智能、计算机视觉和模型轻量化与加速.
E-mail: yangchuanguang@ict.ac.cn

陈路明　男,1991年4月出生,河北石家庄人.93114部队工程师.主要研究方向为机器视觉、深度学习.
E-mail: 295170692@qq.com

赵二虎　男,1985年9月出生,河北邢台人.中国科学院计算技术研究所高级工程师、硕士生导师、特聘研究骨干.主要研究方向为嵌入式智能计算系统.
E-mail: zhaoerhu@ict.ac.cn

安竹林　男,1980年4月出生,山东日照人.中国科学院计算技术研究所副研究员.主要研究方向为人工智能、计算机视觉和模型轻量化与加速.
E-mail: anzhulin@ict.ac.cn

徐勇军　男,1979年7月出生,安徽安庆人.中国科学院计算技术研究所正研级高级工程师、博士生导师、专项技术研究中心主任.主要研究方向为数据智能.
E-mail: xyj@ict.ac.cn